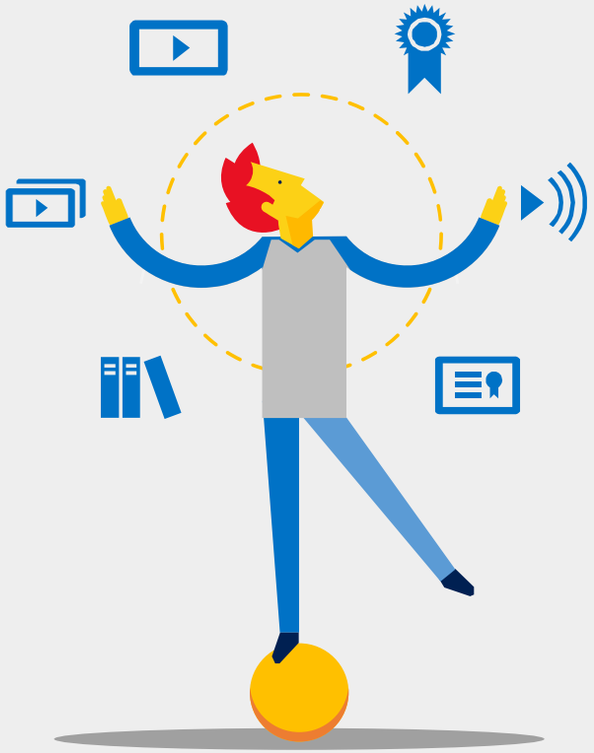
Azure Big Data Landscape

A high level overview of the big data services on the Azure cloud platform

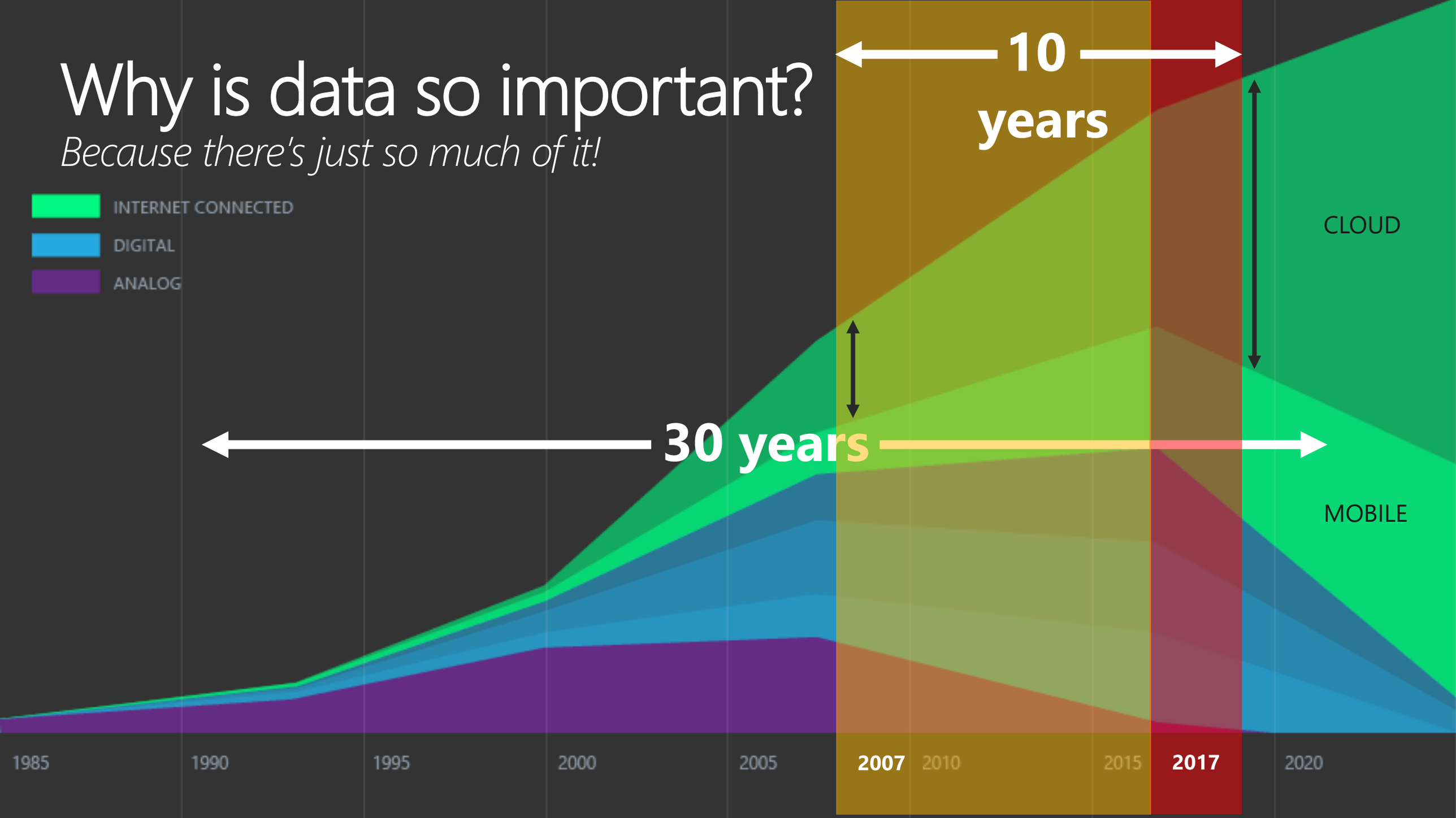
Rolf Tesmer

Azure Data Solutions Architect, Microsoft

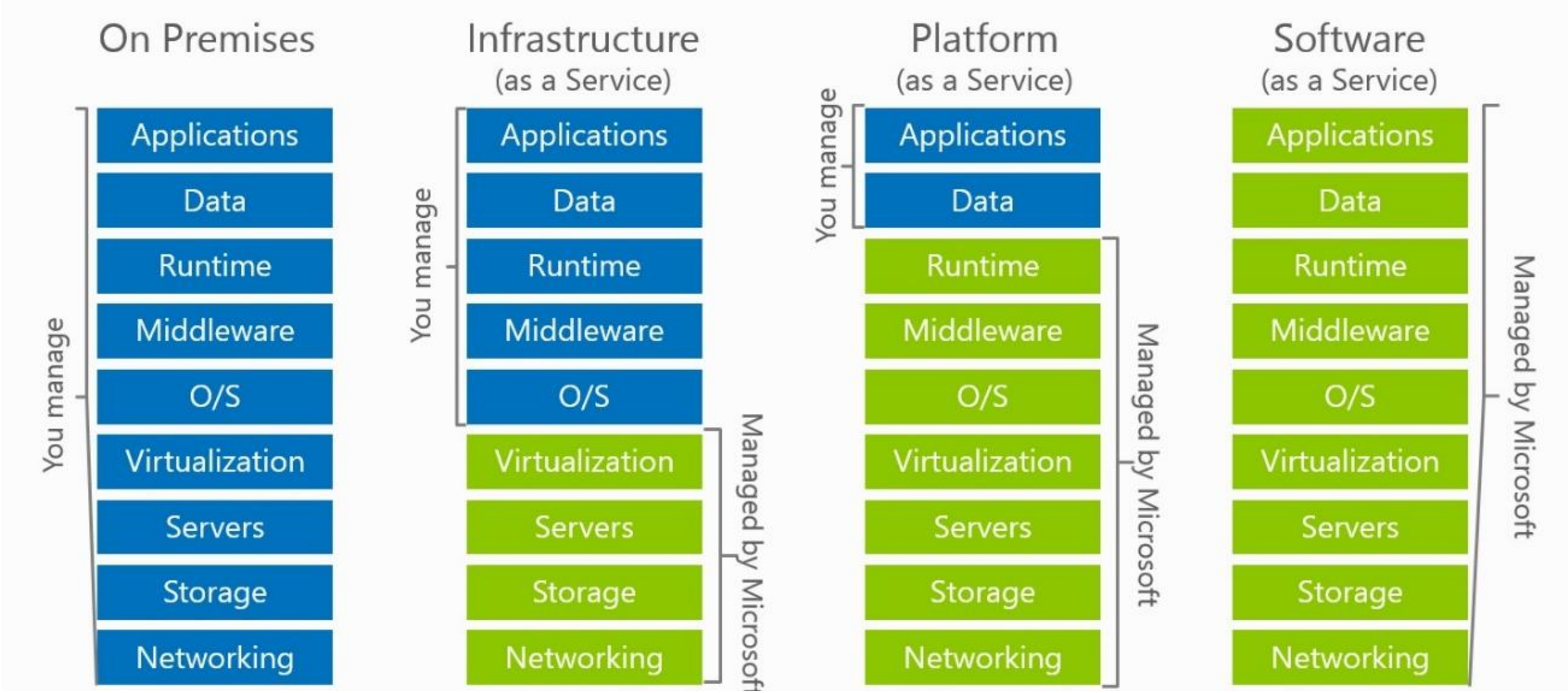


Why is data so important?

Because there's just so much of it!



On-Prem *vs* IaaS *vs* PaaS *vs* SaaS – *Which One?*



http://azureplatform.azurewebsites.net/

* Preview Services

Compute

- Virtual Machines
- Virtual Machine Scale Sets
- Azure Container Service
- Azure Container Registry
- Functions
- Batch
- Service Fabric
- Cloud Services

Networking

- Virtual Network
- Load Balancer
- Application Gateway
- VPN Gateway
- Azure DNS
- Traffic Manager
- ExpressRoute
- Network Watcher

Storage

- Storage: Blobs, Tables, Queues, Files, Disks
- Data Lake Store
- StorSimple
- Azure Backup
- Site Recovery

Web & Mobile

- Web Apps
- Mobile Apps
- Logic Apps
- API Apps
- Content Delivery Network
- Media Services
- Search

Databases

- SQL Database
- SQL Data Warehouse
- SQL Server Stretch Database
- DocumentDB
- Redis Cache
- Data Factory

Intelligence & Analytics

- HDInsight
- Machine Learning
- Cognitive Services
- Azure Bot Service*
- Data Lake Analytics
- Power BI Embedded
- Azure Analysis Services

Internet of Things & Enterprise Integration

- Azure IoT Hub
- Event Hubs
- Stream Analytics
- Notification Hubs
- BizTalk Services
- Service Bus
- Data Catalog

Security + Identity

- Security Center
- Key Vault
- Azure Active Directory
- B2C
- Domain Services
- Multi-Factor Authentication

Developer Services

- Visual Studio Team Services
- Azure DevTest Labs
- VS Application Insights
- API Management
- HockeyApp
- Developer Tools
- Service Profiler*

Monitoring & Management

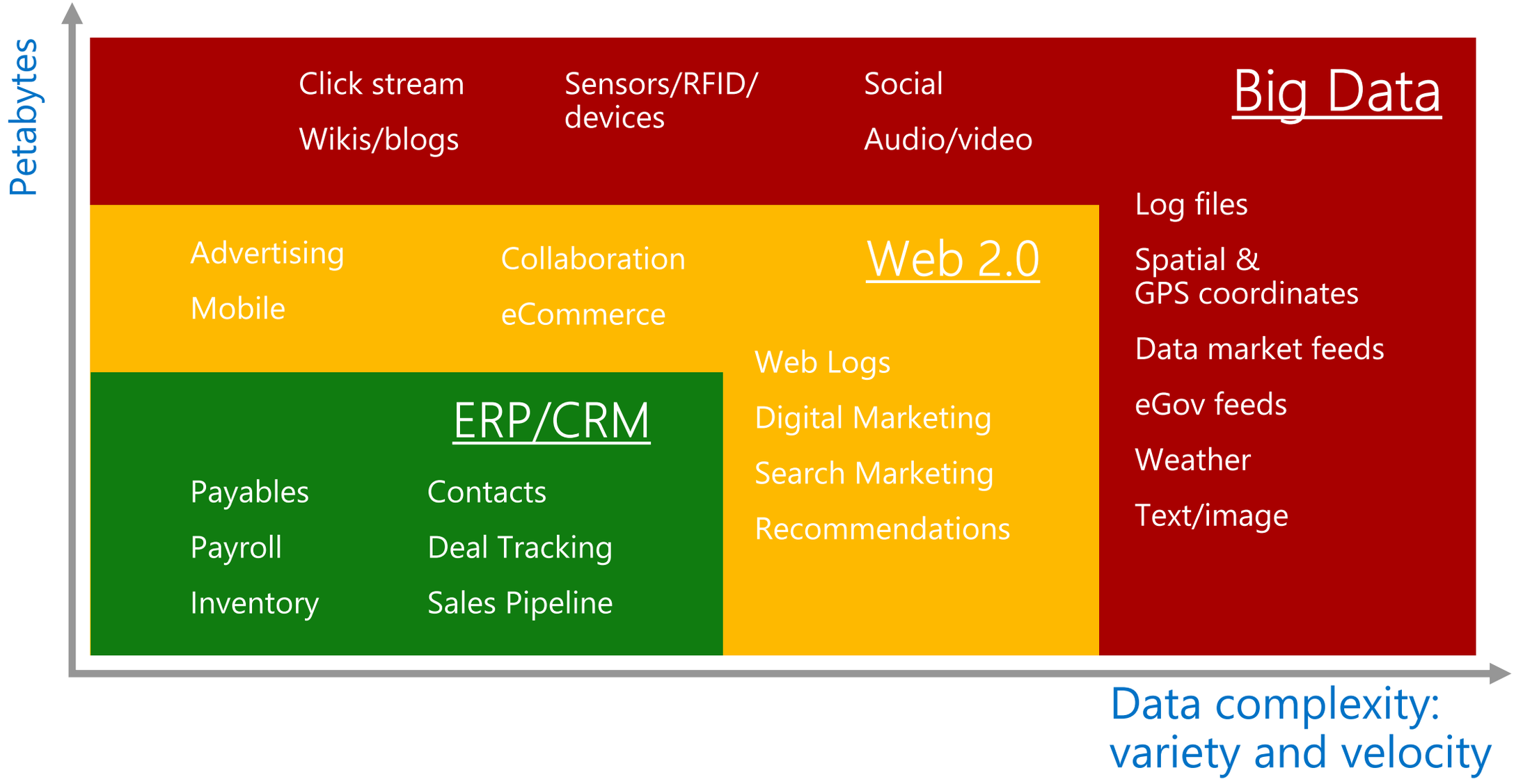
- Azure Portal
- Azure Resource Manager
- Azure Advisor
- Azure Monitor
- Log Analytics
- Automation
- Scheduler

Agenda

2 Key Components of the Microsoft Azure Cloud Data Platform

Big Data / Analytics PaaS
(General Introduction & Overview)

Introduction: Data Size Over the Years…



Introduction: Big Data Definition - *The Four V's*

A **Big Data "***problem***"** exists when you must address **more than one** of the V's.
(**Only one** V indicates current technology is likely to satisfy your goals)

Volume

The data exceeds the physical limits of vertical scalability, implying a scale-out solution (vs. scaling up).

Velocity

The decision window is small compared with the data change rate.

Variety

Many different formats make integration difficult and expensive.

Variability

Many options or variable interpretations confound analysis.

To **solve** the "*problem*" you often **need specialist technologies**
Business wish to **solve** the "*problem*" because it offers **competitive advantage**

Big Data Business Applications & Use Cases

Financial services

New account risk screens
Fraud prevention
Trading risk
Maximize deposit spread
Insurance underwriting
Accelerate loan processing

Retail

360°view of the customer
Analyze brand sentiment
Localized, personalized promotions
Website optimization
Optimal store layout

Telecom

Call detail records (CDRs)
Infrastructure investment
Next product to buy (NPTB)
Real-time bandwidth allocation
New product development

Manufacturing

Supplier consolidation decisions
Supply chain and logistics
Assembly line quality assurance
Proactive maintenance
Crowd source quality assurance

Healthcare

Genomic data for medical trials
Monitor patient vitals
Reduce re-admittance rates
Store medical research data
Recruit cohorts for pharmaceutical trials

Utilities, oil, and gas

Smart meter stream analysis
Slow oil well decline curves
Optimize lease bidding
Compliance reporting
Proactive equipment repair
Seismic image processing

Public sector

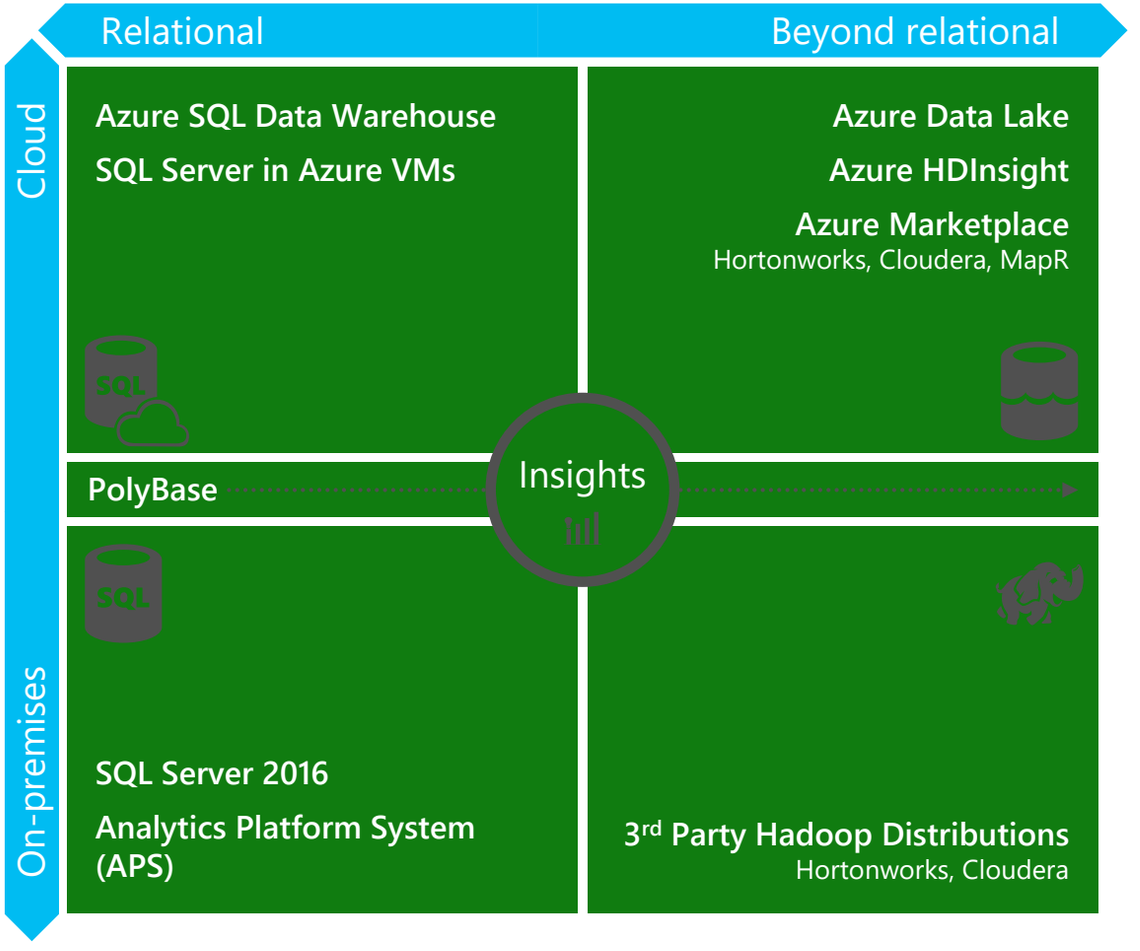
Analyze public sentiment
Protect critical networks
Prevent fraud and waste
Crowd source reporting for repairs to infrastructure
Fulfill open records requests

Big Data and Data Warehousing Compared

Big Data does not negate the business drivers for a **Data Warehouse**. The technologies serve **difference business purposes. Big Data** systems can be a **feeder** into the **Data Warehouse**.

Feature	**Big Data** (ADL, HDInsight, Hadoop, etc)	**Data Warehousing** (SQL DW, SQL in IaaS)
Solution Type	Ecosystem, not a product	Product/Service
Typical Data Type	Structured, Semi-Structured, Unstructured	Structured (Operational)
Typical Data Size	TB – PB Linear Scale out = MPP	GB – TB Non-linear, Scale Up (SMP typically!)
Typical Data Artefacts	Files	Tables/Rows/Columns
Schema	Defined On Read	Defined On Write
Data Consistency, Quality and Accuracy	Low, loose structure, no ACID	High, complex structure, strong ACID
Azure Technologies	HDInsight, Data Lake Vendors (Cloudera, MapR, Hortonworks)	SQL DB, SQL DW SQL Relational Database in IaaS

Big Data as part of a Data Warehousing Solution



Fastest insights

Real-time insights with breakthrough query performance

Analytics built-in

Real-time insights with analytics built in

Choice of deployment

Leading solutions—on-premises and in the cloud

Layers of security

Least vulnerable database 6 years in a row

Any data, any scale

A hybrid solution that grows in step with customer needs

More for the price

Customers do more with industry-leading TCO

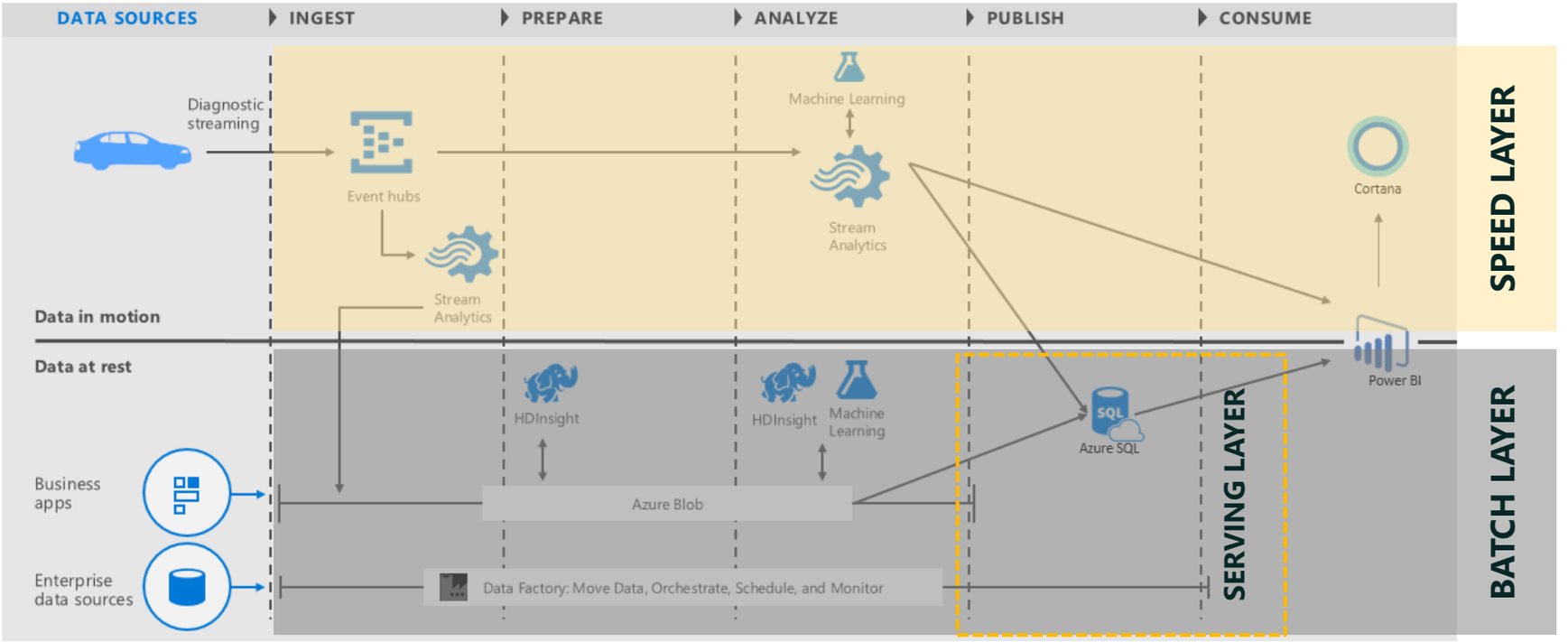
Agenda

1 Lambda Architecture

Big Data / Analytics PaaS

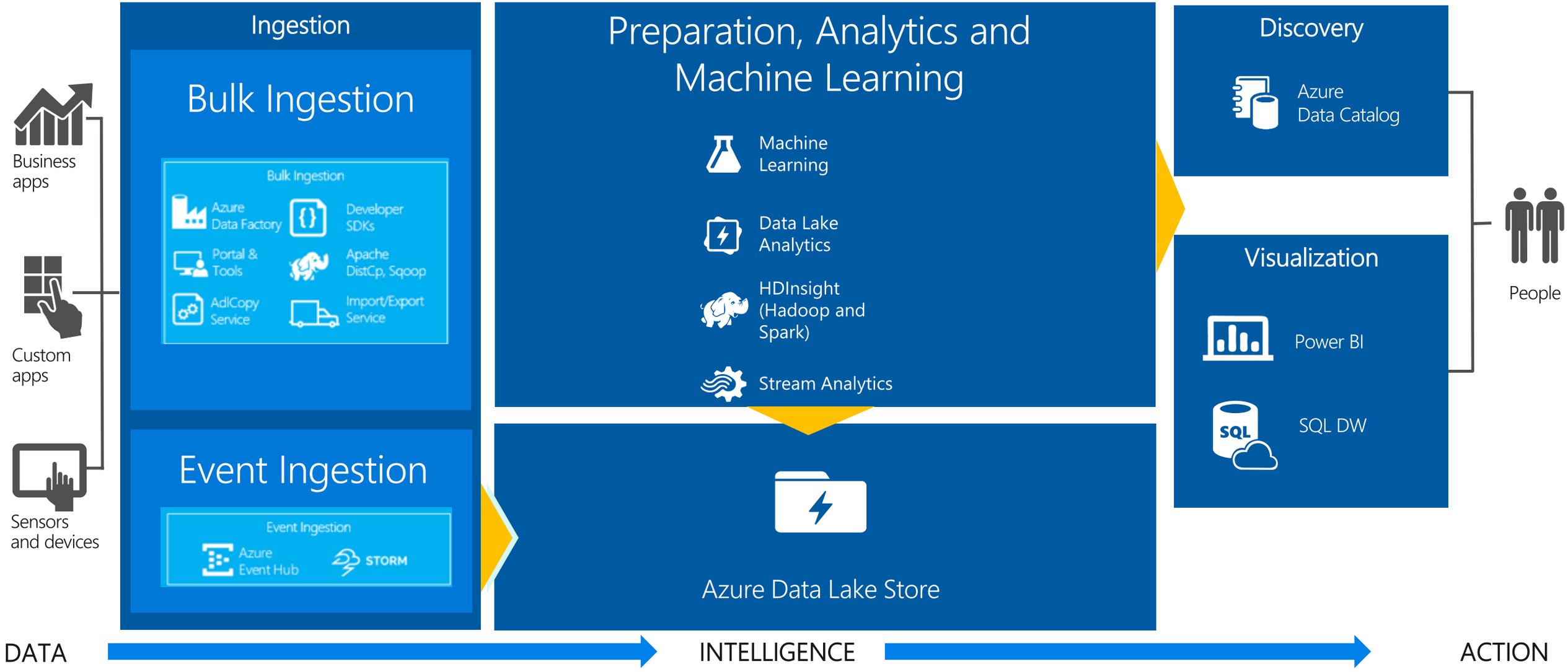
What is the LAMBDA architecture?

*"The Objective of **Lambda Architecture** is to leverage the combined power of both **batch** & **real-time** processing to address the business scenarios where it requires both **historic view of the data** as well as getting insight into the **data in real-time** as business happens."*



https://social.technet.microsoft.com/wiki/contents/articles/33626.lambda-architecture-implementation-using-microsoft-azure.aspx
https://gallery.cortanaintelligence.com/Solution/Telemetry-Analytics
https://docs.microsoft.com/en-us/azure/machine-learning/cortana-analytics-playbook-vehicle-telemetry

Big Data Pipeline and Data Flow in Azure



Agenda

2 What exactly is Unstructured, Semi-Structured and Structured Data?

Big Data / Analytics PaaS
Files, Files and more Files

Considering Data Types

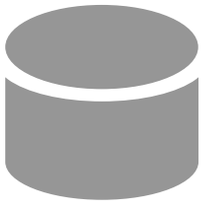
efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk

Unstructured

Store Natively
(logs, pics, etc)

Semi-structured

JSON, XML, etc
→ Schema Evolution (Avro)



Structured

CSV
→ Columnar Storage (Parquet, ORC)

http://www.inquidia.com/news-and-info/hadoop-file-formats-its-not-just-csv-anymore

Columnar Formats: Why? ORC & PARQUET

All data for a column stored contiguously on disk.

So you can read a column really fast.

Just like SQL needs to do.

Pro:

Fast query

Con:

You have to convert data into it

Only do that if you need to query it many times

Columnar Formats: Options

ORCFile:

- Best in Hive
- Allows vectorized execution (Fast)
- Allows ACID (Insert / Update / Delete)

Parquet:

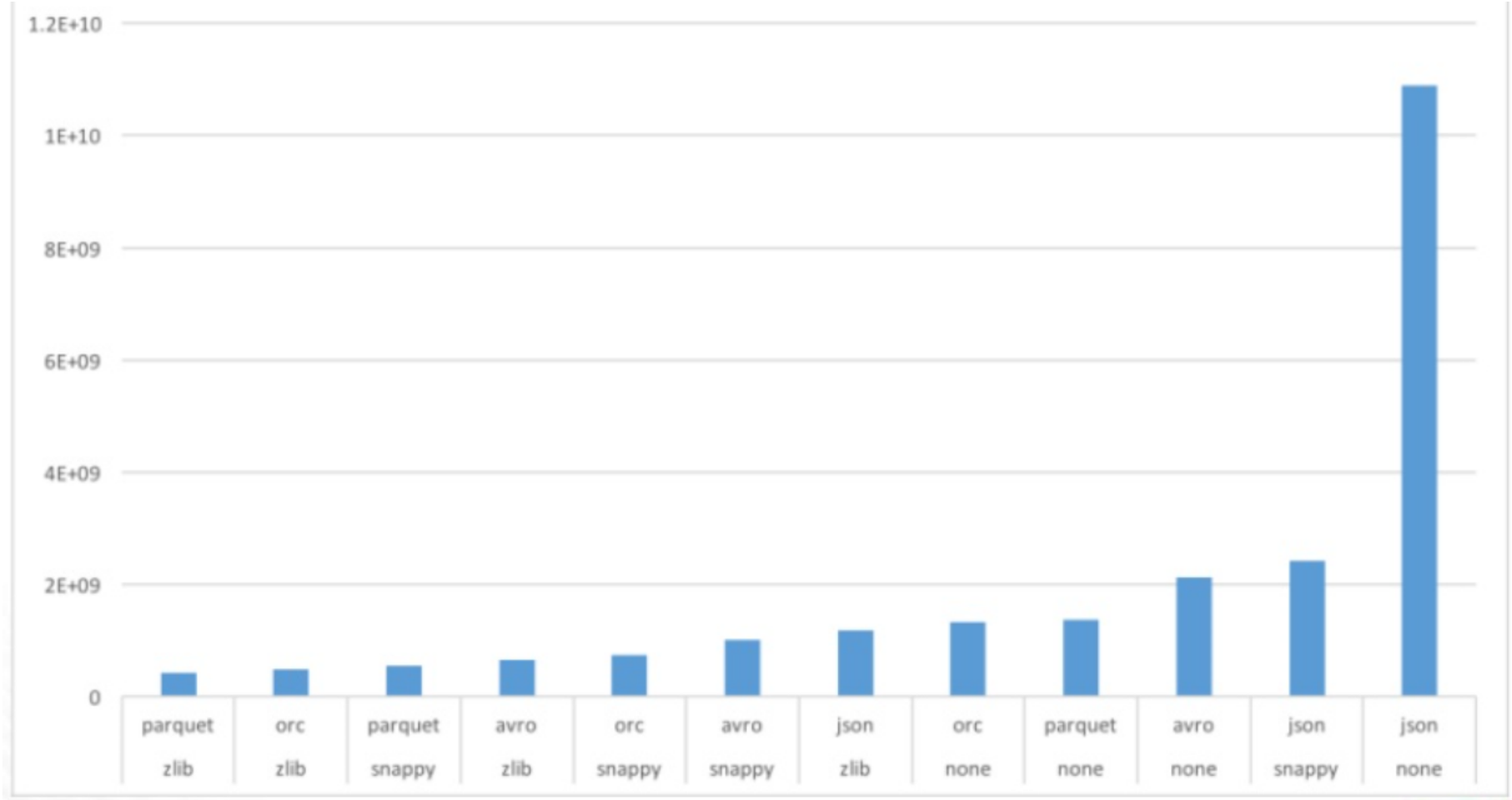
- Fully supported
- No vectorization or ACID
- Common for mixed Hive/Spark workloads

Query Times for Different Formats

Ran 4 queries (using Impala) over 4 Million rows (70GB raw), and 1000 columns (wide table)

Reference: Unknown

Data Size for Different Formats & Compression



Reference: Unknown

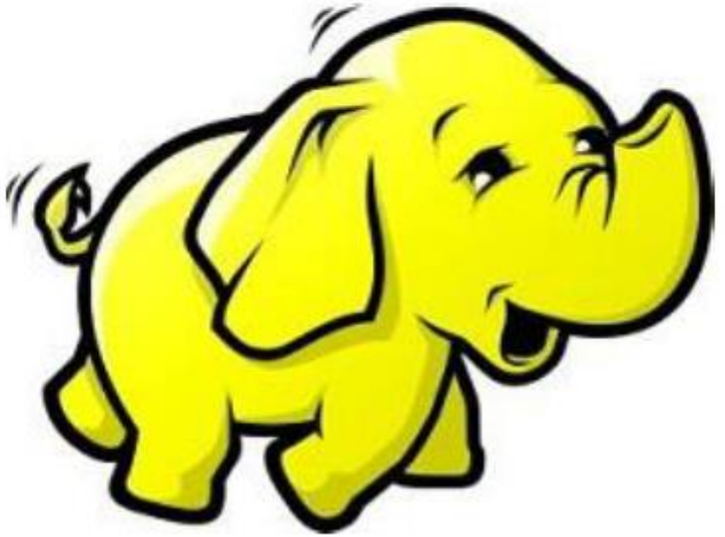
Agenda

2 What exactly is Hadoop?

Big Data / Analytics PaaS
Hadoop

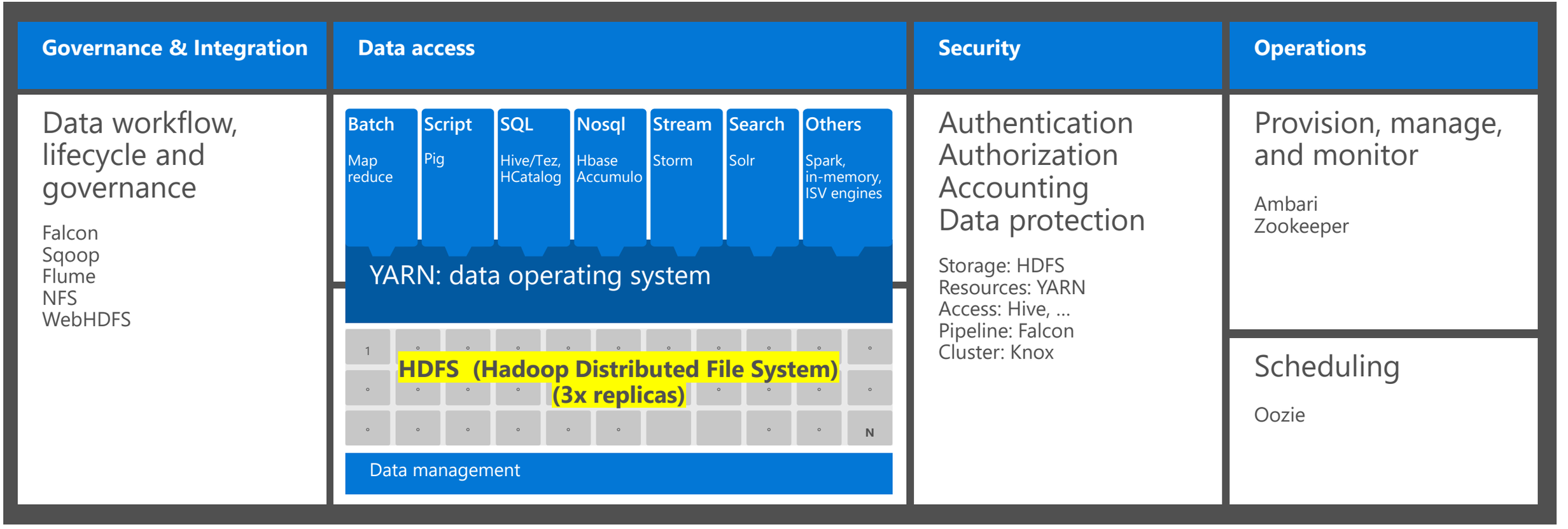
What's the deal with elephants?

"[Hadoop was] the name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid's term" – *Doug Cutting, Hadoop creator*

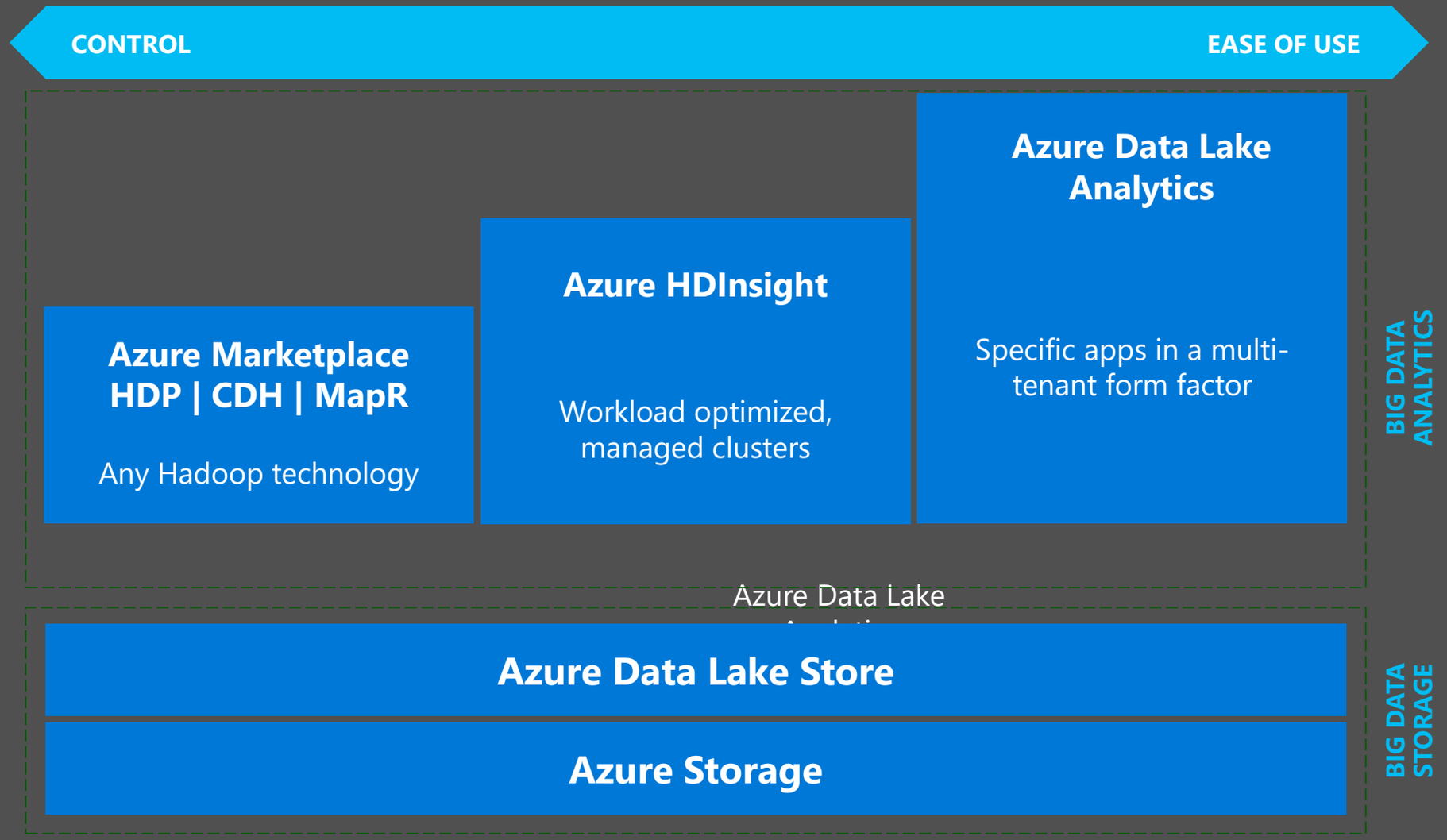


Introduction: What is Hadoop?

A platform with a **portfolio** of **projects**
Governed by Apache Software Foundation (**ASF**) (Open Source)
Comprises core services of **MapReduce**, **HDFS**, and **YARN**



The various big data solutions



Context - Comparing Hadoop and SQL Server

Hadoop		SQL Server
HDFS		Database Windows Cluster
MapReduce YARN Hadoop Common		Relational Engine \SQL OS
Master Web Interface (HUE) Ambari	≈	SQL Server Management Studio
Sqoop		BCP
Hive / Impala		T-SQL (ie Create tables, etc.)
Pig		Powershell
Spark		In Memory SQL Stored Procedures

Reference: "Eating the Elephant" – PASS 2015 - Stuart R Ainsworth

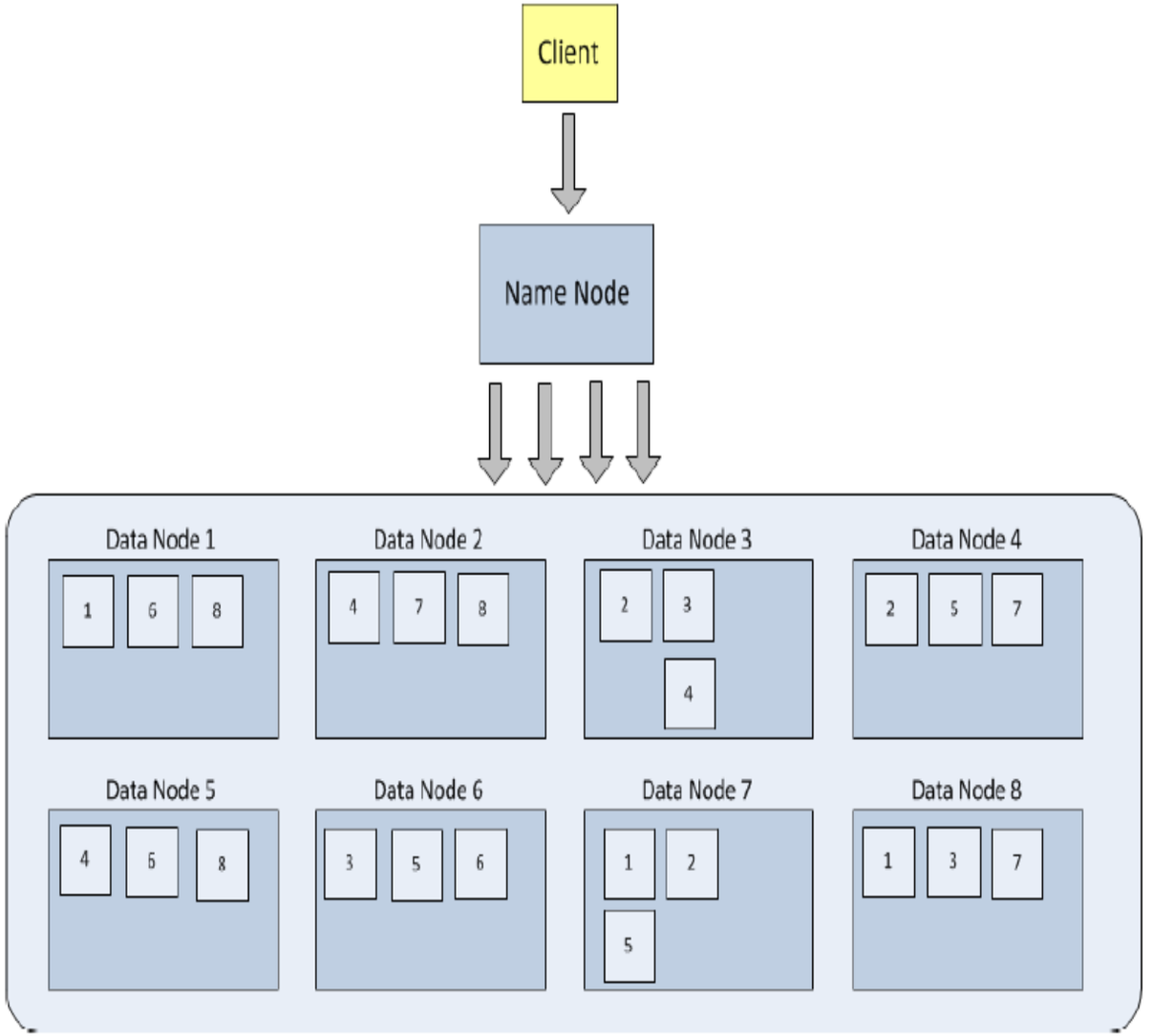
Physical Structure

Data Nodes

- Blocks of data replicated 3 times across Data Nodes
- Store blocks on Local Storage
- Hadoop Distributed File System (HDFS)

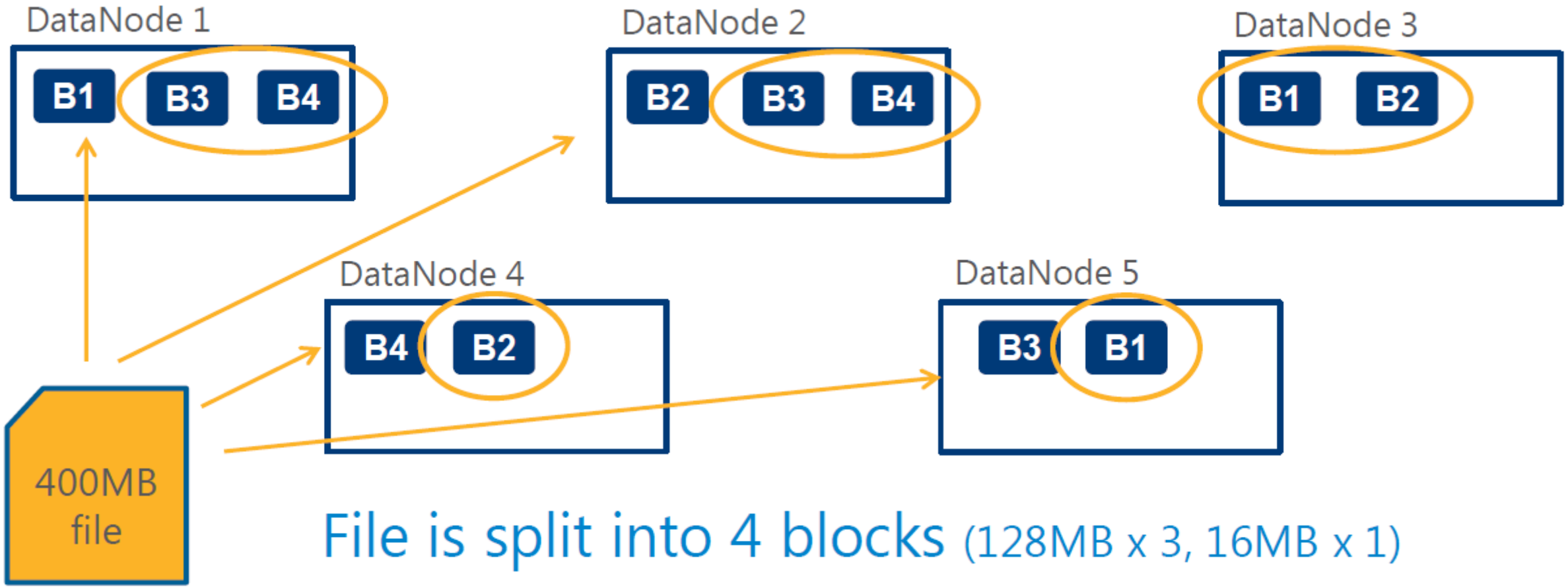
Name Node

- Keeps the directory tree of all files in the file system
- Knows where to get each Block once per request
- Does not store File data itself
- Name node is critical – if down, cluster is down



Data Redundancy

Data Blocks are copied to three different nodes



File is split into 4 blocks (128MB x 3, 16MB x 1)

Agenda

2 Key Components of the Microsoft Azure Cloud Data Platform

Big Data / Analytics PaaS
HDInsight

Azure HDInsight

Hadoop as a Service on Azure (PaaS)

FULLY MANAGED AND **SUPPORTED PaaS**
Hadoop, Spark, Hbase, Storm, Kafka

Available on **LINUX**

100% OPEN SOURCE Apache Hadoop

Clusters up and **RUNNING IN MINUTES (20-30)**

Use familiar **BI TOOLS FOR ANALYSIS** like Excel

HDInsight: Azure PaaS Implementation of Hadoop

HDInsight Supports Several of the Hadoop Projects…

HIVE

- **HiveQL** is a **SQL-like** language (subset of SQL) (Compiled into **MapReduce** jobs)

HBASE

- **Columnar**, **NoSQL** database on data in **HDFS**

SPARK

- **In Memory** Processing on Multiple Workloads

STORM

- **Stream Analytics** for Near-Real Time Processing (similar to Azure Stream Analytics)

HDInsight Supports Hive

SQL-like queries on Hadoop data in HDInsight

HDInsight provides easy-to-use graphical query interface for Hive

HiveQL is a SQL-like language (subset of SQL)

Hive structures include well-understood database concepts such as tables, rows, columns, partitions

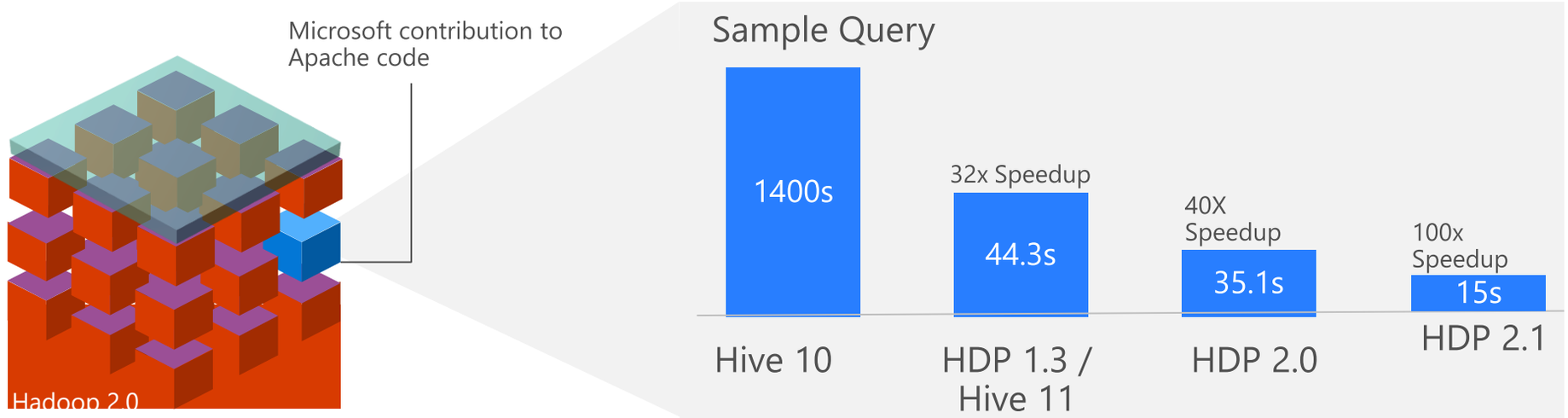
Compiled into MapReduce jobs that are executed on Hadoop

Dramatic performance gains with Stinger/Tez

Stinger is a Microsoft, Hortonworks and OSS driven initiative to bring interactive queries with Hive

Brings query execution engine technology from Microsoft SQL Server to Hive

Performance gains up to 100x



HDInsight Supports Spark

In Memory Processing on Multiple Workloads

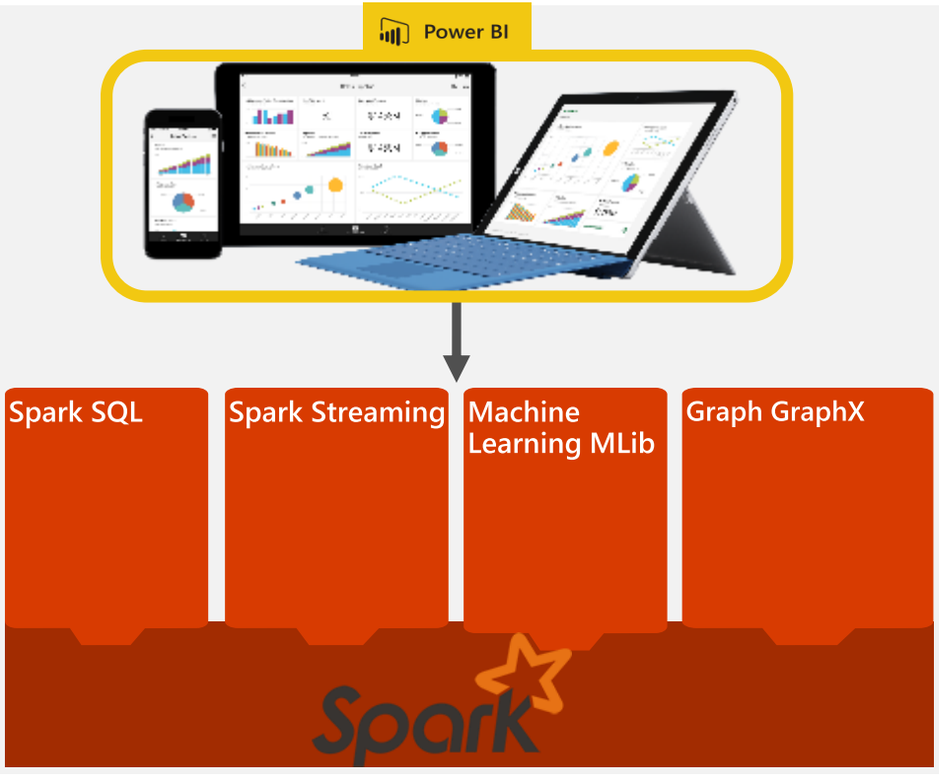
Single execution model for multiple tasks (SQL queries, Streaming, Machine Learning, and Graph)

Processing up to 100x faster performance

Developer friendly (Java, Python, Scala)

BI tool of choice (Power BI, Tabelau, Qlik, SAP)

Notebook experience (Jupyter/iPython, Zeppelin)



Power BI

+ableau

SAP Lumira

QlikView

jupyter

Apache Zeppelin

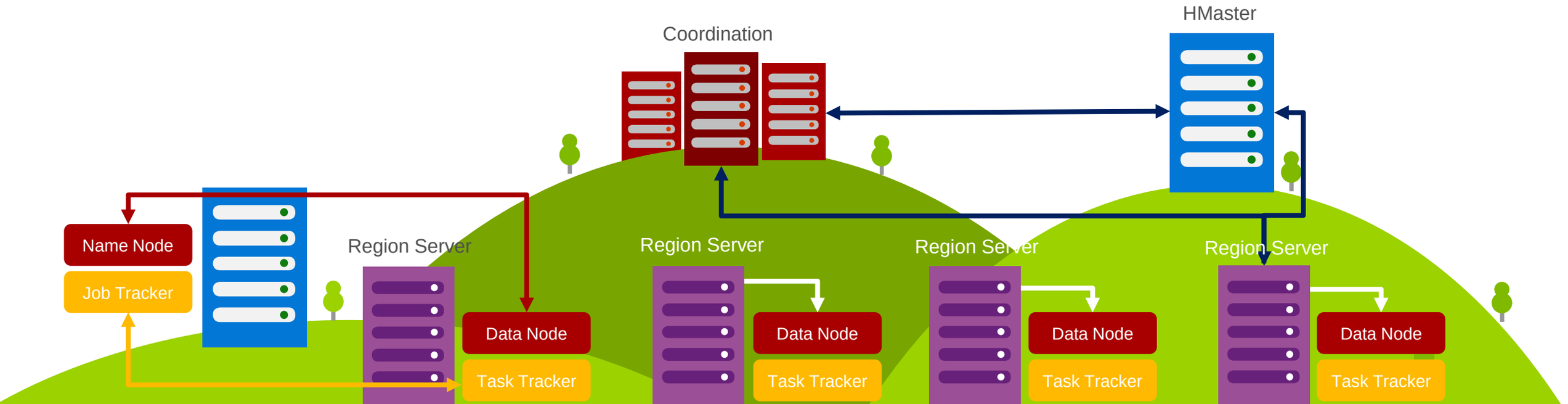
HDInsight Supports HBase

NoSQL database on data in HDInsight

Columnar, NoSQL database

Runs on top of the Hadoop Distributed File System (HDFS)

Provides flexibility in that new columns can be added to column families at any time



HDInsight Supports Storm

Stream analytics for Near-Real Time Processing

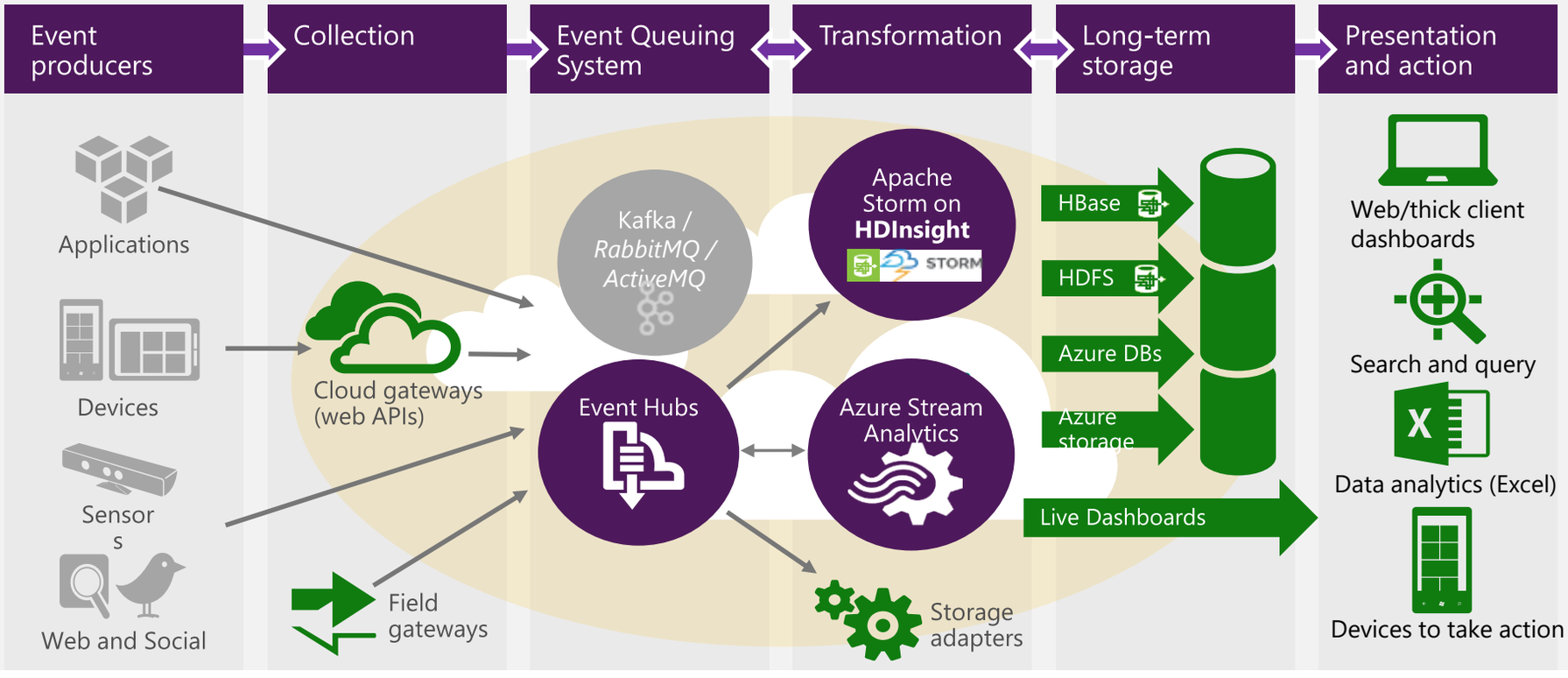
Consumes millions of real-time events from a scalable event broker (ie. Apache Kafka, Azure Event Hub)

Performs time-sensitive computation

Output to persistent stores, dashboards or devices

Customizable with Java + .NET

Deeply integrated to Visual Studio

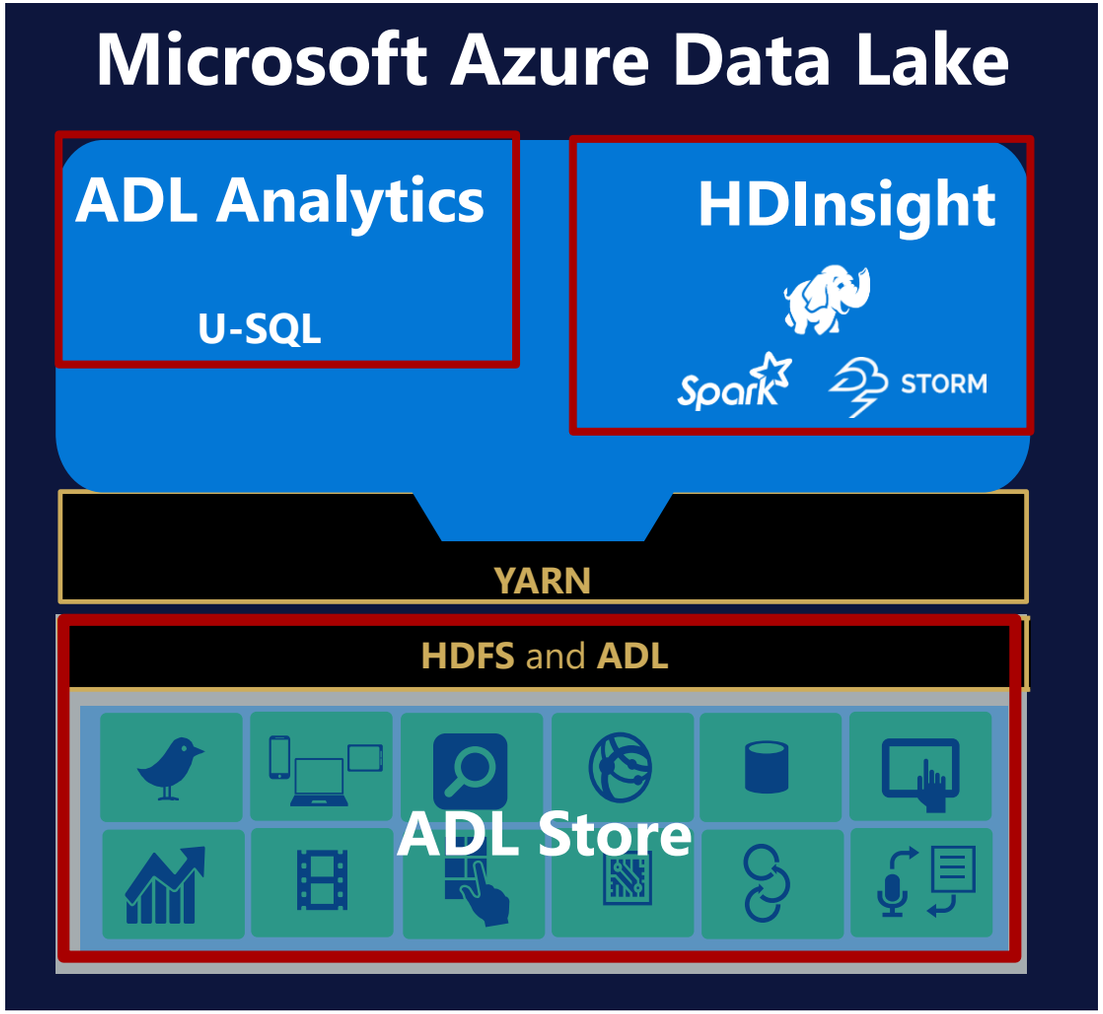


Agenda

2 Key Components of the Microsoft Azure Cloud Data Platform

Big Data / Analytics PaaS
Data Lake Store & Analytics

Introduction: What is Azure Data Lake Store & Analytics?



Consists of 2 component parts;
Data Lake Store & Data Lake Analytics

Distributed PaaS service

Both Instantly scale to meet performance needs

Analytics over **all data**
(unstructured, semi-structured, structured)

U-SQL to perform Analytics
(simple and familiar, easily extensible)
(Integrated into **Visual Studio** tools)

Built on open standards (**YARN**)

Can deploy other services on store (ie HDInsight)

Azure Data Lake Store

A hyper scale repository for big data analytics workloads

An enterprise wide repository of every type of data collected in a single place prior to any formal definition of requirements or schema.

SCALE No limits

ANY DATA Store in its native format

HADOOP FILE SYSTEM (HDFS) for the cloud

NATIVELY accessible via both HDFS and ADL

ENTERPRISE READY access control, encryption

PERFORMANCE Optimized for analytic workload

PaaS Service managed by Microsoft

Azure Data Lake Store – Technical Details

Durable & Highly Available

- Data is managed by Microsoft (PaaS)

Unlimited Storage

- **Unlimited** account sizes, **no limits** to scale
- Individual file sizes to **PBs**

Secure

- Secure files and folders, **POSIX** (ACL)
- **Auditing** and **logging**
- **Encryption** at rest

Optimised for Analytic Workloads

- Designed for large scale **parallel** processing
- Auto optimize to match **active workloads**
- **Immediate** read after write

Primary Use Cases

- Long term IoT storage
- Clickstream analysis
- Social analysis
- Web log analysis
- File based batch processing
- Staging files for DW loads
- Long term DW archive
- (+ *similar use cases to Big Data*)

Azure Data Lake Analytics

A elastic analytics service built on Apache YARN that processes all data, at any size

AUTO SCALE with no limits

U-SQL a language that unifies the benefits of SQL with the expressive power of C#

Optimized to work with **ADL STORE**

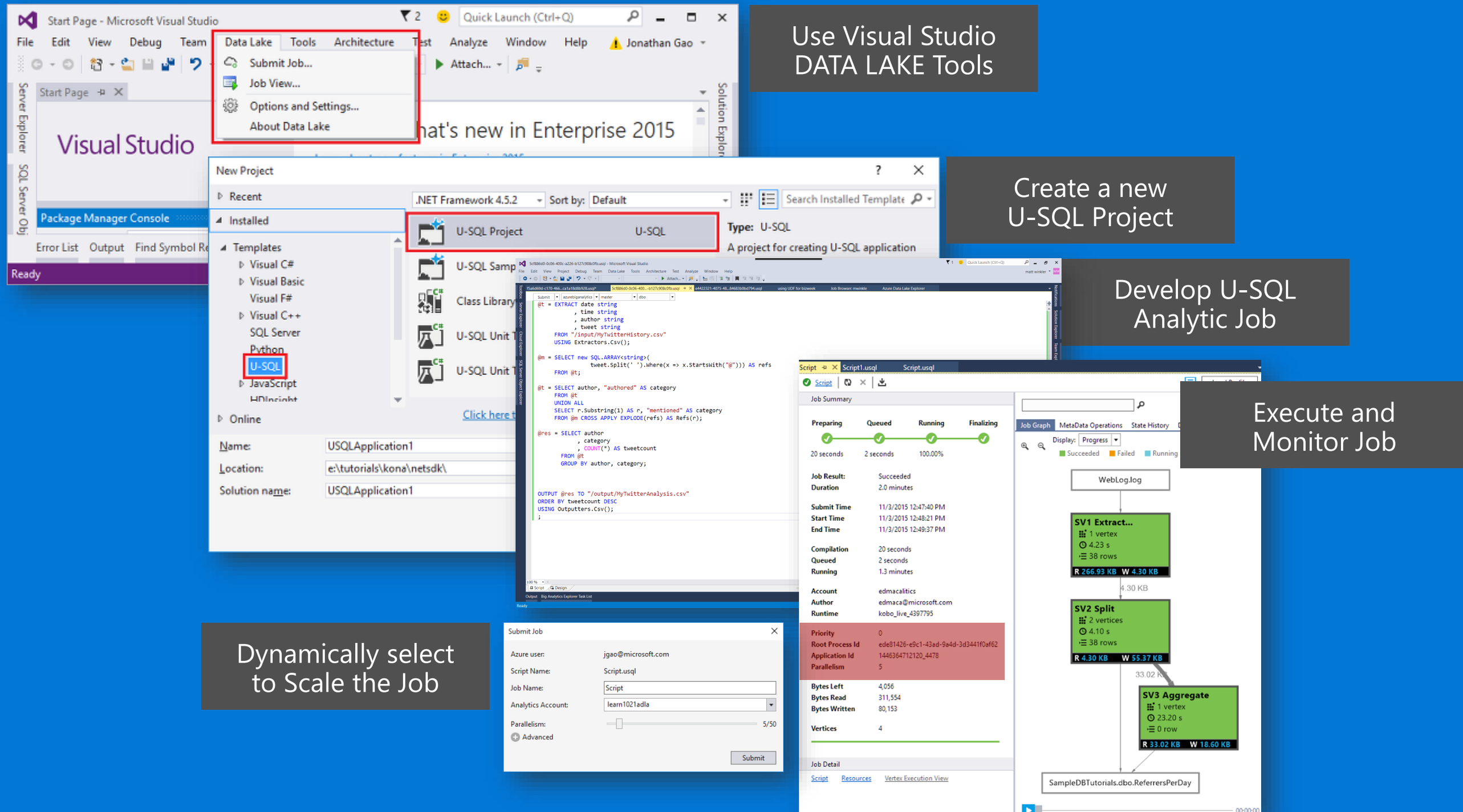
FEDERATED QUERY with Azure data sources

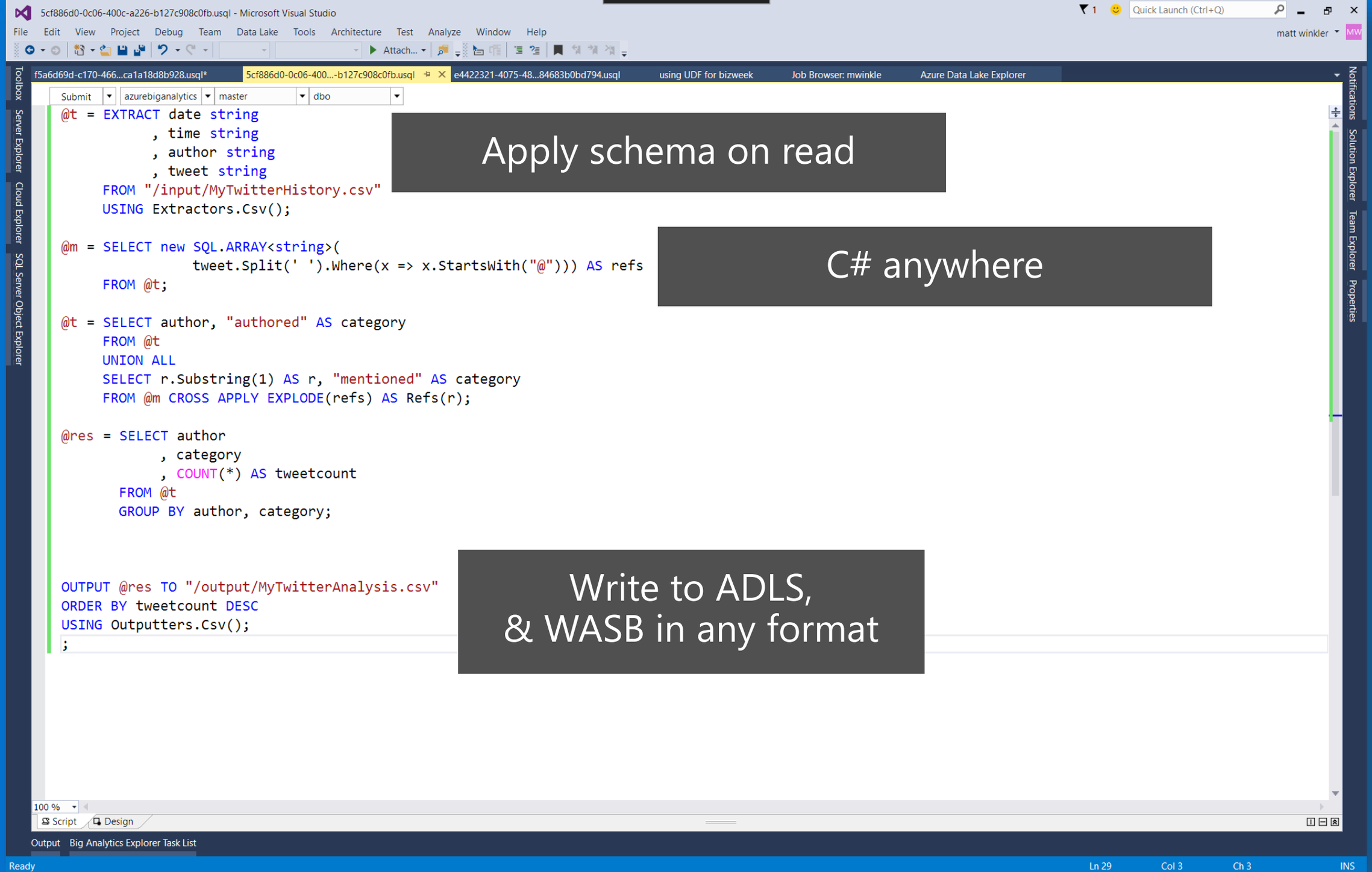
ENTERPRISE READY

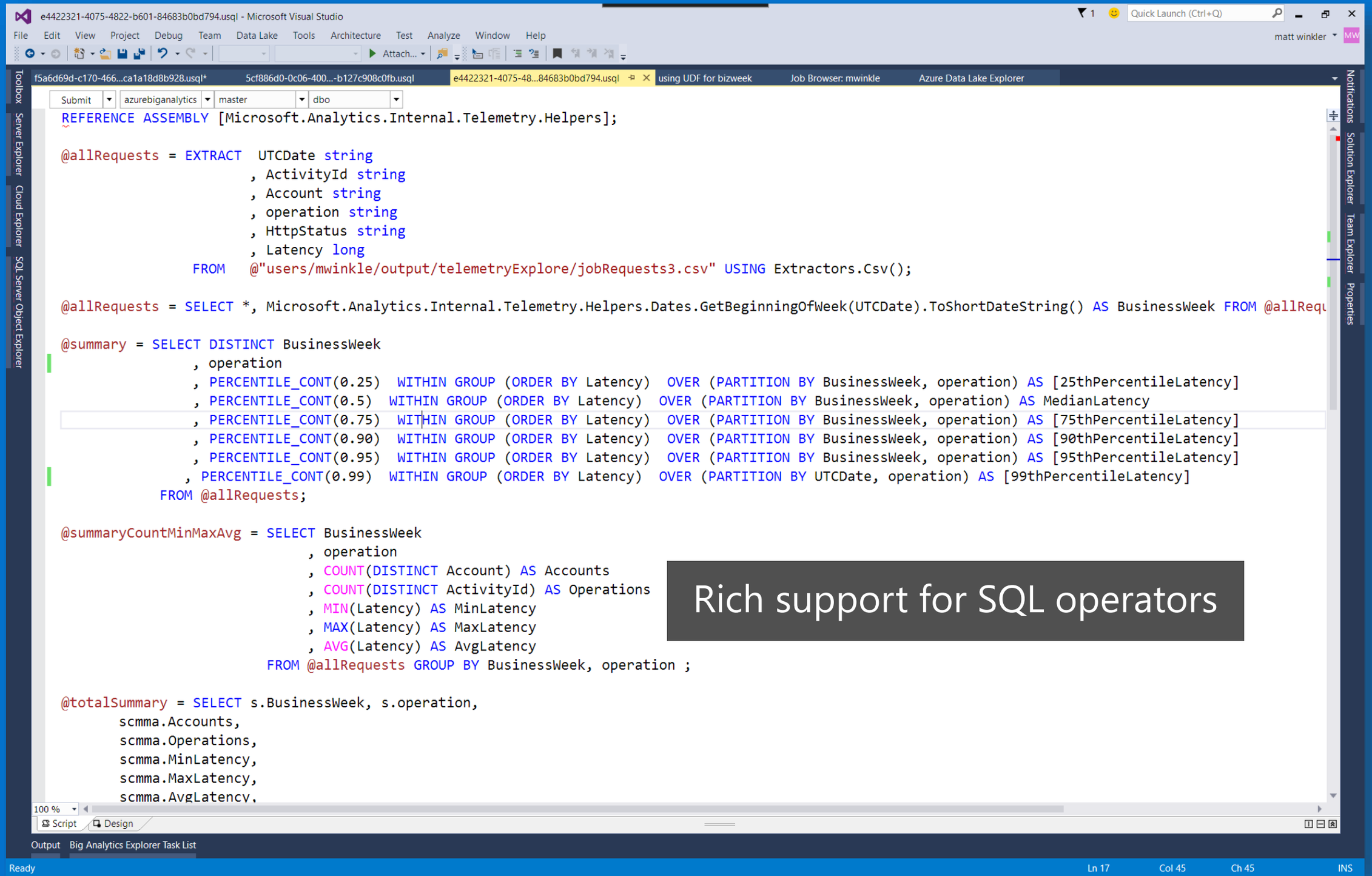
Pay & Auto Scale **PER (U-SQL) ANALYTIC JOB**

DEVELOP jobs in **Visual Studio** or **Azure Portal**

U-SQL Reference: https://msdn.microsoft.com/en-us/library/azure/mt591959.aspx
Example Code: https://blogs.msdn.microsoft.com/robinlester/2016/01/04/an-introduction-to-u-sql-in-azure-data-lake/


Use Visual Studio DATA LAKE Tools
Start Page - Microsoft Visual Studio
Data Lake
Submit Job...
Job View...
Options and Settings...
About Data Lake
Create a new U-SQL Project
New Project
U-SQL Project
Develop U-SQL Analytic Job
Execute and Monitor Job
Dynamically select to Scale the Job
Submit Job
Azure user: jgao@microsoft.com
Script Name: Script.usql
Job Name: Script
Analytics Account: learn1021adla
Parallelism: 5/50
Submit
WebLog.log
SV1 Extract...
SV2 Split
SV3 Aggregate
SampleDBTutorials.dbo.ReferrersPerDay

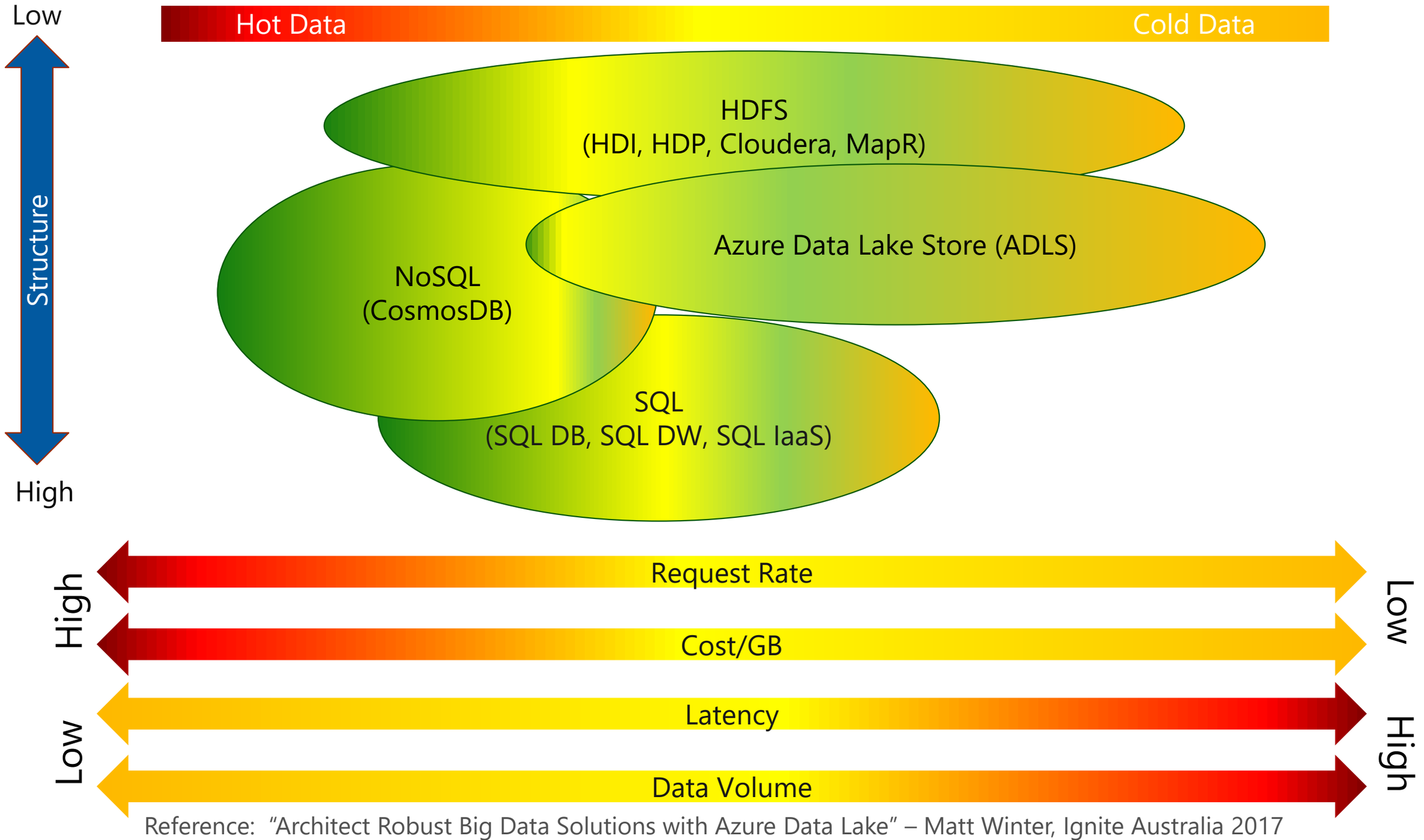

5cf886d0-0c06-400c-a226-b127c908c0fb.usql - Microsoft Visual Studio
File Edit View Project Debug Team Data Lake Tools Architecture Test Analyze Window Help
matt winkler
Submit azurebiganalytics master dbo
@t = EXTRACT date string
, time string
, author string
, tweet string
FROM "/input/MyTwitterHistory.csv"
USING Extractors.Csv();
@m = SELECT new SQL.ARRAY<string>(
tweet.Split(' ').Where(x => x.StartsWith("@"))) AS refs
FROM @t;
@t = SELECT author, "authored" AS category
FROM @t
UNION ALL
SELECT r.Substring(1) AS r, "mentioned" AS category
FROM @m CROSS APPLY EXPLODE(refs) AS Refs(r);
@res = SELECT author
, category
, COUNT(*) AS tweetcount
FROM @t
GROUP BY author, category;
OUTPUT @res TO "/output/MyTwitterAnalysis.csv"
ORDER BY tweetcount DESC
USING Outputters.Csv();
;
Apply schema on read
C# anywhere
Write to ADLS,
& WASB in any format
Script Design
Output Big Analytics Explorer Task List
Ready
Ln 29 Col 3 Ch 3 INS


e4422321-4075-4822-b601-84683b0bd794.usql - Microsoft Visual Studio
File Edit View Project Debug Team Data Lake Tools Architecture Test Analyze Window Help
matt winkler
Submit azurebiganalytics master dbo
REFERENCE ASSEMBLY [Microsoft.Analytics.Internal.Telemetry.Helpers];
@allRequests = EXTRACT UTCDate string
, ActivityId string
, Account string
, operation string
, HttpStatus string
, Latency long
FROM @"users/mwinkle/output/telemetryExplore/jobRequests3.csv" USING Extractors.Csv();
@allRequests = SELECT *, Microsoft.Analytics.Internal.Telemetry.Helpers.Dates.GetBeginningOfWeek(UTCDate).ToShortDateString() AS BusinessWeek FROM @allRequ
@summary = SELECT DISTINCT BusinessWeek
, operation
, PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [25thPercentileLatency]
, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS MedianLatency
, PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [75thPercentileLatency]
, PERCENTILE_CONT(0.90) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [90thPercentileLatency]
, PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [95thPercentileLatency]
, PERCENTILE_CONT(0.99) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY UTCDate, operation) AS [99thPercentileLatency]
FROM @allRequests;
@summaryCountMinMaxAvg = SELECT BusinessWeek
, operation
, COUNT(DISTINCT Account) AS Accounts
, COUNT(DISTINCT ActivityId) AS Operations
, MIN(Latency) AS MinLatency
, MAX(Latency) AS MaxLatency
, AVG(Latency) AS AvgLatency
FROM @allRequests GROUP BY BusinessWeek, operation ;
@totalSummary = SELECT s.BusinessWeek, s.operation,
scmma.Accounts,
scmma.Operations,
scmma.MinLatency,
scmma.MaxLatency,
scmma.AvgLatency,
Rich support for SQL operators
Output Big Analytics Explorer Task List
Ready
Ln 17 Col 45 Ch 45 INS

Agenda

2 Comparing How the Technologies Overlap

Big Data



Reference: "Architect Robust Big Data Solutions with Azure Data Lake" – Matt Winter, Ignite Australia 2017

Data Processing Technology Choices

	Azure SQL DW	Azure HDInsight with Spark	Cloudera Impala	Azure HDInsight (Hive, Pig, etc.)	Azure Data Lake Analytics
Query Latency	Low	**Low**	Low	**Medium (Tez),** High (MapReduce)	High
Durability	High	High	High	High	High
Data Volume	Up to 60 TB	*nodes	*nodes	*nodes	***vertices**
Managed	Yes	Yes	**No**	Yes	Yes
Storage	SQL DB	Blob, ADLS	HDFS	ADLS, Blob	ADLS, Blob
SQL Compatibility	**High**	Low (SparkSQL)	Medium	Medium (HiveQL)	Medium

Reference: "Architect Robust Big Data Solutions with Azure Data Lake" – Matt Winter, Ignite Australia 2017

Agenda

3 Q & A

References

- Big Data - https://msdn.microsoft.com/en-us/library/dn749868.aspx
- Hadoop - https://en.wikipedia.org/wiki/Apache_Hadoop
- Map Reduce - https://en.wikipedia.org/wiki/MapReduce
- Hive - https://en.wikipedia.org/wiki/Apache_Hive
- Spark (core, streaming, ML, graphX) - https://en.wikipedia.org/wiki/Apache_Spark
- Storm - https://en.wikipedia.org/wiki/Storm_(event_processor)
- Kafka - https://en.wikipedia.org/wiki/Apache_Kafka
- Sqoop - https://en.wikipedia.org/wiki/Sqoop
- Impala - https://en.wikipedia.org/wiki/Cloudera_Impala
- Cloudera - https://en.wikipedia.org/wiki/Cloudera
- Hortonworks - https://en.wikipedia.org/wiki/Hortonworks
- Data Lake - https://en.wikipedia.org/wiki/Hortonworks
- HDInsight - https://msdn.microsoft.com/en-us/library/dn749853.aspx
- Mahout - https://en.wikipedia.org/wiki/Apache_Mahout
- Avro - https://wiki.apache.org/hadoop/Avro/
- Parquet - https://en.wikipedia.org/wiki/Apache_Parquet
- ORC - https://orc.apache.org/docs/
- SPARK vs IMPALA – which and when - https://learning.naukri.com/articles/spark-vs-impala/
- Patterns and Practices - https://msdn.microsoft.com/en-us/library/dn749804.aspx